PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 :
G06F 17/30, H04L 29/02, 12/56, H01J 3/02

A2

(11) International Publication Number: **WO 98/57275**

(43) International Publication Date: 17 December 1998 (17.12.98)

(21) International Application Number: PCT/SE98/01085

(22) International Filing Date: 8 June 1998 (08.06.98)

(30) Priority Data:
9702239-6 12 June 1997 (12.06.97) SE

(71) Applicant *(for all designated States except US)*: TELIA AB [SE/SE]; Mårbackagatan 11, S-123 86 Farsta (SE).

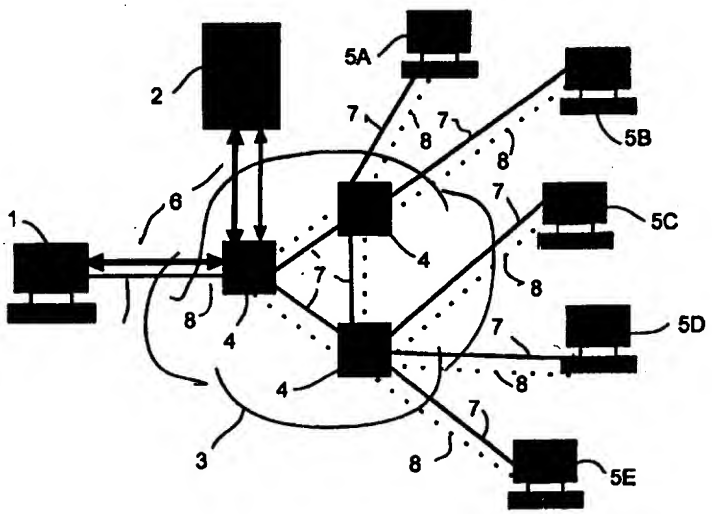
(72) Inventor; and
(75) Inventor/Applicant *(for US only)*: KAVAK, Nail [SE/SE]; Myrstugevägen 359, S-143 32 Vårby (SE).

(74) Agent: PRAGSTEN, Rolf; Telia Research AB, Vitsandsgatan 9, S-123 86 Farsta (SE).

(81) Designated States: EE, JP, LT, LV, NO, US, European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

Published
Without international search report and to be republished upon receipt of that report.

(54) Title: ARRANGEMENT FOR LOAD SHARING IN COMPUTER NETWORKS



(57) Abstract

The invention relates to an arrangement for load sharing in computer networks and, more exactly, an arrangement for distribution of traffic, for instance via Internet, from clients (1) to service suppliers who provide services from many servers. The invention makes possible distribution to one of a number of replicated servers. Suitable server is selected for instance on basis of available resources at the interface of the server, or less delay in the transmission. The invention results in better performance and reduced traffic by distribution of the traffic geographically and from a resource point of view. According to the invention, a number of replicated servers (5A–5E) belong to an anycast–group and each anycast–group is connected to a domain name server (2) which has the ability to select one of the replicated servers, so that a router (4) can establish a connection between the selected server and the service–requesting client computer. Each replicated server (5A–5E) can transmit a resource advertisement which contains information about available resources at the server in question, and about the link parameters of the server.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania
AM	Armenia
AT	Austria
AU	Australia
AZ	Azerbaijan
BA	Bosnia and Herzegovina
BB	Barbados
BE	Belgium
BF	Burkina Faso
BG	Bulgaria
BJ	Benin
BR	Brazil
BY	Belarus
CA	Canada
CF	Central African Republic
CG	Congo
CH	Switzerland
CI	Côte d'Ivoire
CM	Cameroon
CN	China
CU	Cuba
CZ	Czech Republic
DE	Germany
DK	Denmark
EE	Estonia
ES	Spain
FI	Finland
FR	France
GA	Gabon
GB	United Kingdom
GE	Georgia
GH	Ghana
GN	Guinea
GR	Greece
HU	Hungary
IE	Ireland
IL	Israel
IS	Iceland
IT	Italy
JP	Japan
KE	Kenya
KG	Kyrgyzstan
KP	Democratic People's Republic of Korea
KR	Republic of Korea
KZ	Kazakstan
LC	Saint Lucia
LI	Liechtenstein
LK	Sri Lanka
LR	Liberia
LS	Lesotho
LT	Lithuania
LU	Luxembourg
LV	Latvia
MC	Monaco
MD	Republic of Moldova
MG	Madagascar
MK	The former Yugoslav Republic of Macedonia
ML	Mali
MN	Mongolia
MR	Mauritania
MW	Malawi
MX	Mexico
NE	Niger
NL	Netherlands
NO	Norway
NZ	New Zealand
PL	Poland
PT	Portugal
RO	Romania
RU	Russian Federation
SD	Sudan
SE	Sweden
SG	Singapore
SI	Slovenia
SK	Slovakia
SN	Senegal
SZ	Swaziland
TD	Chad
TG	Togo
TJ	Tajikistan
TM	Turkmenistan
TR	Turkey
TT	Trinidad and Tobago
UA	Ukraine
UG	Uganda
US	United States of America
UZ	Uzbekistan
VN	Viet Nam
YU	Yugoslavia
ZW	Zimbabwe

TITLE OF THE INVENTION: ARRANGEMENT FOR LOAD SHARING IN COMPUTER NETWORKS

FIELD OF THE INVENTION

The present invention relates to an arrangement for load sharing in computer networks, and more exactly, an arrangement for distribution of traffic, for instance via Internet, from clients to the service suppliers who provide services from a multiple of servers. The invention makes possible distribution to a number of replicated servers. Suitable server is selected, for instance, on basis of available resources at the server's interface or on less delay in the connection. The invention results in better performance and reduced traffic by distribution of the traffic geographically and from a resource point of view.

PRIOR ART

Many internationally big companies provide a multiple of copies of their information servers, in some cases also abroad, with the aim to increase accessability and performance for the Internet user. One problem is that some servers always are more loaded than others, especially because audio- and video file transmissions in real time are becoming more and more popular. This results in congestions in the networks.

A conventional way of solving this problem is to replace an existing server by another which has higher process capacity and storage capacity. In most cases, however, the problem is not lack of capacity but lack of load sharing, i.e. the solution often is to utilize the existing network resources better.

In some cases the information is copied or replicated in a multiple of servers which are distributed geographically in different places in order to improve response times. Traditionally, the traffic is distributed on the different servers by using a sequential (round

robin) or random technology. In this sequential technology the servers are selected sequentially in turn, whereas random technology selects servers in just any order. In none of these technologies information about the load on the servers, or of the localization of the servers in the network, is utilized.

The present invention solves the above mentioned problems by utilizing information about the load and the network topology of the servers, together with a technology to address suitable servers, namely the anycast-technology (anycast routing and anycast addressing). By that, the invention can distribute traffic to geographically distributed servers so that the traffic is distributed to for instance less loaded server.

SUMMARY OF THE INVENTION

Consequently the present invention provides an arrangement for load sharing in a computer network, comprising a larger number of users or clients with computers, at least one service supplier which provides services via a number of replicated servers, a computer network with routers which can connect the client computers with the servers in order to connect a client computer which requests a service with suitable server.

According to the invention a number of replicated servers belong to a common address group (anycast-group) and each anycast-group is connected to a domain name server which has the ability to select one of the replicated servers, so that a router can establish a connection between the selected server and the service-requesting client computer.

Preferably the domain name server is arranged to select the least loaded replicated server or the nearest replicated server. Each replicated server can transmit a resource message which contains information about available

resources at the server in question, and about the link parameters of the server.

Thanks to the new technology according to the present invention, a number of advantages are achieved. The end users make benefit from the lower delay and higher performance. The number of servers in the network or with the service supplier can be reduced. It will be easy to add or subtract servers without the users becoming interfered with, i.e. without interrupting services in progress. This implies considerably reduced investment costs and managing costs for the network operator and the service supplier, while the end user at the same time is offered higher performance. The end users are not aware of that there are several servers which provide the same service.

The invention provides cost efficient solutions which can be configured automatically and are transparent to the user. The invention makes possible step by step expansion of servers and adapts the capacity according to the need. By routing traffic to the least loaded server, the number of jumps and potential time delays are reduced.

BRIEF DESCRIPTION OF THE DRAWINGS

A preferred embodiment of the invention will be described in detail below with reference to enclosed drawings where:

Figure 1 is a diagram over the arrangement according to the invention, and

Figure 2 is an illustration over the format of a resource message according to the invention.

DETAILED DESCRIPTION OF PREFERRED EMBODIMENTS

One of the more important limited resources for instance in World Wide Web is network bandwidth and processor capacity of servers and client computers. The network bandwidth and the processor power are increasing

from a general point of view, but not fast enough to keep pace with the more and more increasing number of users in the network. The number of users soon will be a billion.

This would be no problem if the users restricted their activities to their local machines, but the network makes possible for the users to request documents from servers in distant located places. It is also a fact that some servers are more popular than others, which means that these servers have to carry a too heavy load in spite of that there are a lot of servers which can deliver the same services. A solution to this problem is to store documents nearer the users. This will reduce both the network traffic and the load of the servers. A similar solution is a replication or copying which essentially implies buffer storing before the documents are requested. The cost of multireplication can be reduced by using flow transmission distribution as is supported by NNTP (Network News Transfer Protocol). Further, replication can be applied both at services and documents. If replicated servers are in clusters near the original server, this will reduce the load on the single server, but does not reduce the network traffic because all requests are still going to the cluster. If, however, replicated servers are distributed over the network, and if clients automatically can locate the nearest or least loaded server, this would reduce the network traffic as well.

The aim of the present invention is to provide an arrangement by which the network resources can be used more efficiently. The aim is to dynamically route the user traffic to a least loaded application server. The advantages which at that are achieved, are mentioned above. Even if the invention is described with special reference to IP over ATM (i.e. networks which operate with Internet-protocol with Asynchronous Transfer Mode), the invention can as well be applied to other technologies than ATM.

In Figure 1 is shown a diagram over a preferred arrangement according to the invention. A client computer 1 has a virtual circuit connection established to its domain name servers (DNS) 2, of which one is shown in the figure, via the network 3. There may be a secondary connection to a reserve domain name server (not shown) which can be accessed at error or at break/interruption. In the network there are a number of routers 4, which establish connections with servers 5A to 5E at service suppliers'. All the servers 5A, 5B, 5C, 5D, 5E are copies of each other and constitute the same anycast-group or "common address group". Anycast-groups only transmit a virtual IP-address to their neighbours. A logical address can correspond to a number of physical addresses.

The domain name server has knowledge of all anycast-groups. For scaling- and performance reasons anycast-groups can be distributed over many domain name servers. The anycast-group members are connected to the domain name servers by a point-to-multipoint-connection.

The function of the arrangement can be summarized as follows. The client computer first transmits a request to a domain name server to resolve a domain name into an IP-address. Control messages to and from the domain name server are represented by the bold arrows 6 in the figure. The domain name server first controls the semantics of the DNS-request in order to find out which application that was requested. For instance, an ftp-request is handled in a different way than a ping-request. The former is long-lived, whereas the latter is short-lived. The type of application assists the domain name server in selecting the most relevant server for each request. In some cases the least loaded server is selected, whereas in other cases the cheapest link and belonging server is selected.

In an IPATM-network a transmitter sometimes prefers to use the shortest route for time critical applications, for instance pictures, video, ftp, and in other cases the same

transmitter would prefer the cheapest route, such as ping, dns etc.

The domain name server receives continuously routing information from all anycast-servers which it is serving. The routing information is transmitted via the connections marked 7 in the Figure 1. The routing information includes details about the available capacity for each link, the number of jumps to each server, the bandwidth of each link, the processor capacity, measured delay etc. The domain name server returns the IP-address to the server which is best adapted to the application demands which are derived from the DNS-request. When the IP-address has been resolved, it is returned to the client computer so that a direct connection is established between the client computer and the server. The connection is transparent to the domain name server. The server or document which is requested, then can be transferred to the client computer. The data flow is shown at 8 in Figure 1.

To select one server out of the different suitable, exactly alike servers, anycast-technology is used. An anycast-address (common address) is used to represent a group of nodes, one of which shall be selected. When an anycast-address is received, the domain name server delivers the IP-address to all destinations which are represented by the anycast-address. Since many servers can have the same anycast-address, the router conventionally selects one of them by collecting information about the number of jumps to the different servers and selects the nearest one. Use of links is another at present available criterion. Other possible criteria are costs, processor load, storage, routing policy etc.

One important aspect of the invention is to use the anycast-technology to access services which are delivered by a larger number of servers which are operating on different network nodes. Especially the same service can be accessed by one single anycast-address irrespective of how

many servers that are used for the service and of where the servers are located in the network. To copy the service at many servers over the network is useful when high service accessibility is required. Other services which might benefit from anycast-technolgy include: World Wide Web, video services, domain name servers, address resolving servers, Neighbour Discovery-servers, 020-number services, telephoning via Internet, and the Public Switched Telephone Network in order to find the most cost efficient voice services etc.

An anycast-server, ANS, which can be located together with the domain name server 2, holds information about the membership for all anycast-service members 5A-5E. A node can register itself at the anycast-service, join as a member, withdraw from and stop participating in the ANS-service.

To make load sharing distributed to least loaded server working, each server need to inform the surroundings abouts its accessible resources. To this purpose a Resource Advertisement (RA) is transmitted. At calculation of service quality routes (QoS routes), RA is used.

Each server produces a resource advertisement for each coverage area and indicates the largest amount of accessible resources for reservations on the interface of each of the servers in the coverage area, together with the delay parameters of the link. These parameters are roughly analogous to the standard cost value "Open Shortest Path First" (OSPF), but are independent of the standard service type value to better characterize the static delay characteristics of a link. A new copy of the resource message is produced whenever a new routing resource advertisement is produced for the coverage area, or whenever the available bandwidth resource or the delay is changed for a link in the coverage area. An algorithm can be used so that a new resource advertisement is produced only when the available bandwith resource is changed to a

considerable extent. Resource advertisements are transmitted in flows through one single coverage area. The format of the resource advertisement is shown i Figure 2.

The number of links is the link number which is included in the resource advertisement. For each link, the link type, link identity, and link data, are the same as for the routing resource advertisement. The available link resource is represented by packet data area parameters in floating point format with simple precision according to IEEE, as in the service model Control Load. The link delay is a statistical delay value for the link expressed in milliseconds.

Input to the routing calculation includes source address and destination address and the service qualities for the flow, which at present are the packet data area parameters from the Path-message of the resource reservation protocol, but might also be derived from other trigger units. In order to calculate the best route, or the route with least delay, the statistical delay value is used in the same way as OSPF uses the zero cost value of the TOS (Type of Service) of the router.

The DNS-server need both information about available resources and existing resource reservation in additon to the normal information about topology and membership.

When a new replicated server connects to an anycast-group, the local router informs about the existence of the service as a part of its normal routing exchanges with the neighbouring routers. If the local router observes that a replicated server has stopped transmitting messages, also this negative information is transmitted to the neighbouring routers. Each of the neighbouring routers studies a distance value for the announced anycast-server, updates its tables based on this value, and forwards in its turn the updated information to its neighbours. Each router maintains knowledge at least of the route to the nearest anycast-servers and possibly a small list over alternative

servers in case it is informed that the previous nearest server is down. In this way every router in the network will have the sufficient knowledge to route anycast-packets to the nearest server, but need not maintain a database over all anycast-servers in the network. If a client transmits a request for connection to the anycast-group, and if it is not a service in the local subnetwork, the local router forwards the packet to the nearest replicated server based on the current routing tables.

It is assumed that an anycast-call can be routed to one of the servers on a normal level in the server hierarchy, irrespective of where these servers are phycially located in the network. For many applications the normal level would be either the level for the anycast-caller's node, or nearest level above. The anycast-caller shall not need to know any details about how the service hierarchy is organized, or even how the levels in the server hierarchy are numbered. All that the caller need to specify is that an anycast-call should be routed to a server on the lowest possible higher level where a server can be found. To meet the demands and address a server hierarchy by one single address, the invention suggests to widen the definition of proximity which is used by routing algorithms to include new parameters to measure proximity. Suppose that PNNI-routing levels (Private Network Node Interface) are used to identify the hierarchical levels of the servers. If we let the difference between two routing level identifiers be an acceptable measure of proximity between two corresponding nodes, an anycast-caller need not directly specify a level in the hierarchy. By use of this measurement parameter routing to servers on the lowest possible higher level in a hierarchy of servers can be supported, irrespective of the physical distance between the requested server and the client.

Consequently the present invention provides an arrangement for load sharing which solves the indicated

problems. An expert in the field realizes that the invention can be implemented in a number of different ways with different combinations of hardware and software without leaving the frame of the invention. The extent of protection of the invention is only limited by the patent claims below.

PATENT CLAIMS

1. Arrangements for load sharing in computer networks containing

a larger number of client computers (1),

at least one service supplier which provides services via a number of replicated servers (5A-5E).

a computer network (3) with routers (4) which can connect the client computers (1) with the servers (5A-5E) to connect a client computer which requests a service with suitable server, c h a r a c t e r i z e d in that a multiple of replicated servers (5A-5E) belong to an anycast-group.

that each anycast-group is connected to a domain name server (2) which has the ability to select one of the replicated servers (5A-5E), so that a router (4) can establish a connection between the selected server and the service requesting client computer (1).

2. Arrangement according to patent claim 1, c h a r a c t e r i z e d in that the domain name server (2) is arranged to select the least loaded replicated server (5A-5E).

3. Arrangement according to patent claim 2, c h a r a c t e r i z e d in that each replicated server (5A-5E) is arranged to transmit a resource advertisement (RA) which contains information about available resources of the server.

4. Arrangement according to patent claim 1, c h a r a c t e r i z e d in that the domain name server (2) is arranged to select the nearest replicated server (5A-5E).

5. Arrangement according to patent claim 4, c h a r a c t e r i z e d in that each replicated server (5A-5E) is arranged to transmit a resource advertisement (RA) which contains information about the link parameters

of the server, for instance link delay parameters or routing level.

6. Arrangement according to any of the previous patent claims, c h a r a c t e r i z e d in that the domain name server (2) is arranged to utilize anycast-technology to select a replicated server (5A-5E).

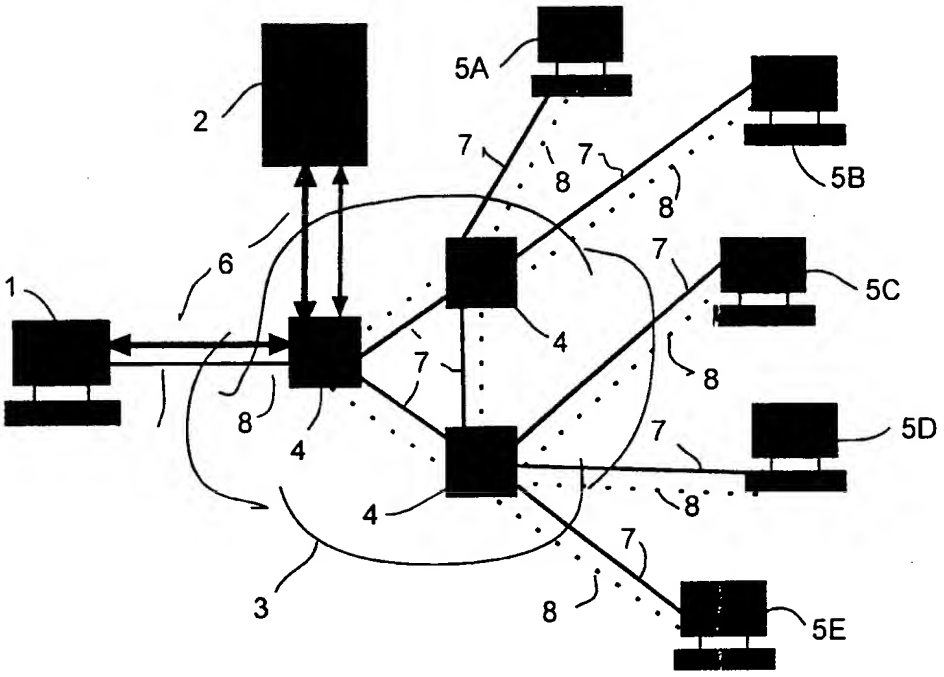


Figure 1

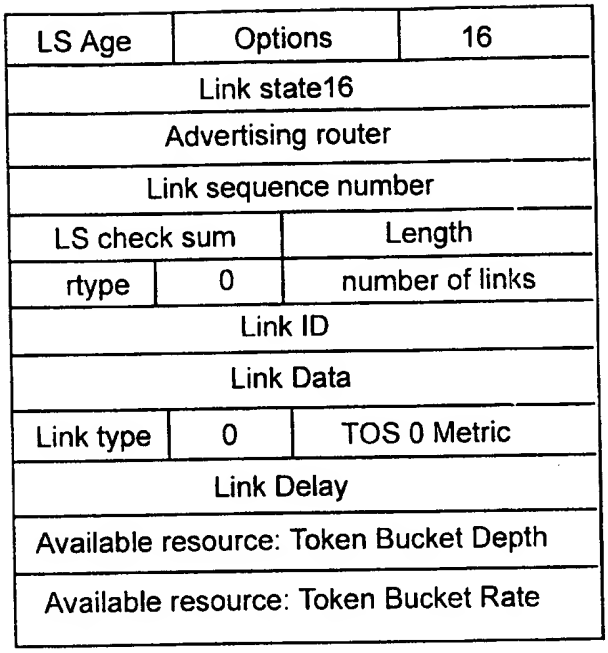

LS Age
Options
16
Link state16
Advertising router
Link sequence number
LS check sum
Length
rtype
0
number of links
Link ID
Link Data
Link type
0
TOS 0 Metric
Link Delay
Available resource: Token Bucket Depth
Available resource: Token Bucket Rate

Figure 2